AD-A285 867

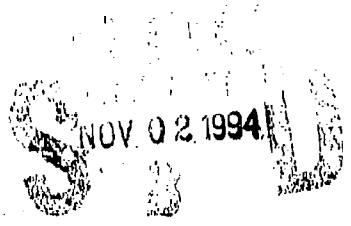
(1)

ARMY RESEARCH LABORATORY

Latin Hypercube Sampling in Sensitivity Analysis

Joseph C. Collins, III

ARL-TR-586 October 1994



DTIC QUALITY INSPECTED 3

94-33972

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION IS UNLIMITED.

94 11 1 108

NOTICES

Destroy this report when it is no longer needed. DO NOT return it to the originator.

Secondary distribution of this report is prohibited.

The findings of this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

The use of trade names or manufacturers' names in this report does not constitute indorsement of any commercial products.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1994	Final: Oct 92-May 94

4. TITLE AND SUBTITLE

Latin Hypercube Sampling in Sensitivity Analysis

5. FUNDING NUMBERS

PR: 1L162618AH80

6. AUTHOR(S)

Joseph C. Collins, III

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

U.S. Army Research Laboratory
ATTN: AMSRL-SL-BV
Aberdeen Proving Ground, MD 21005-5068

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Research Laboratory
ATTN: AMSRL-OP-AP-L
Aberdeen Proving Ground, MD 21005-5066

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

ARL-TR-586

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

In practice, production computer simulation models are usually run for a limited number of "real" input vectors. To make statistical inferences about the model, it is frequently desirable to generate additional "realistic" inputs for the purpose of obtaining better coverage of the input space. Latin Hypercube Sampling is presented as a method for accomplishing this task.

14. SUBJECT TERMS

experimental design, matrices (mathematics)

15. NUMBER OF PAGES

28

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Intentionally Left Blank.

Acknowledgements

The author thanks reviewers Wendy Winner and Andrew Thompson for providing numerous insightful comments. Incorporating their suggestions has improved the readability of the report. Any further shortcomings remain the responsibility of the author.

A-1

Intentionally Left Blank.

Contents

Intentionally Left Blank.

Figures

Tables

Intentionally Left Blank.

1. Introduction

1.1. Background

In the course of vulnerability modeling or any other computer simulation activity, the need arises to quantify certain aspects of the behavior of the model itself as an autonomous system. This is distinct from use of the model as a predictive or analytic tool, in that the simulation model itself is now the subject of study. In the words of Iman[1] *et. al.*:

> "... it is important to have efficient techniques to examine and assess the influence of model input on model output. That is, it is important to be able to perform sensitivity analyses on the relationship between information supplied to the model and predictions made by the model. The benefits of such analyses include the following: 1) an indication whether the model operates as intended, 2) identification of unimportant variables or unnecessary model complexity, and 3) an assessment of relative input variable importance for guidance in data collection."

A directive to analyze the behavior of the compartment model[2] for the purpose of determining the relative importance of its input variables had led to the application of methodology presented in this report. The techniques presented herein are applicable to the sensitivity analysis of computer simulations in general, and consideration should be given to their incorporation into such analyses.

1.2. The Basic Problem

Conceptually, the vector input $\mathbf{x}$ and scalar output y of a simulation model are functionally related by $y = F(\mathbf{x})$ where the function is unknown. Of interest is the local sensitivity of the model (*i.e.*, the relationship between changes in $\mathbf{x}$ and changes in y when $\mathbf{x}$ is centered about a single fixed operating point with input $\mathbf{x}_0$ and output $y_0 = F(\mathbf{x}_0)$).

The local Taylor series representation of F at $\mathbf{x}_0$ is

$$F(\mathbf{x}) = F(\mathbf{x}_0) + \frac{d}{d\mathbf{x}}F(\mathbf{x}_0)^t\Delta\mathbf{x} + \Delta\mathbf{x}^t\frac{d^2}{d\mathbf{x}^2}F(\mathbf{x}_0)\Delta\mathbf{x} + \ldots \qquad [1]$$

where $\Delta\mathbf{x} = \mathbf{x} - \mathbf{x}_0$ is the incremental change in $\mathbf{x}$ about the operating point. Vectors are columns, and throughout this presentation $\mathbf{A}^t$ denotes the transpose of $\mathbf{A}$. Truncating the Taylor series to first order, we obtain the approximation

$$F(\mathbf{x}) = F(\mathbf{x}_0) + \frac{d}{d\mathbf{x}}F(\mathbf{x}_0)^t\Delta\mathbf{x} \qquad [2]$$

or

$$\Delta y = \mathbf{b}^t\Delta\mathbf{x} \qquad [3]$$

where

$$\Delta y = y - y_0 = F(\mathbf{x}) - F(\mathbf{x}_0) \qquad [4]$$

is the incremental change in y about the operating point y_0 and the derivative vector $\mathbf{b} = \frac{d}{d\mathbf{x}} F(\mathbf{x}_0)$ relates Δy to $\Delta \mathbf{x}$. The components of **b** thus quantify the sensitivity of the model to changes in input and allow us to answer questions about the relative importance of the various input dimensions.

Let us suppose now that we have a number of observations $(\Delta \mathbf{x}_i, \Delta y_i)$, each representing a slight variation in the model input and output about the operating point. We may construct a vector **Y** with component Y_i equal to Δy_i and a matrix **X** with row i equal to $\Delta \mathbf{x}_i^t$. Noting that $\mathbf{b}^t \Delta \mathbf{x}_i = \Delta \mathbf{x}_i^t \mathbf{b}$, the collection of equations [3] may be written succinctly as

$$\mathbf{Y} = \mathbf{X}\mathbf{b} \qquad [5]$$

which expresses the problem of estimating **b** in the language of linear regression.

1.3. Design Considerations

The problem here is to estimate the tangent plane of a multidimensional surface at a single point. We assume that the true response is a "nice" function (*i.e.*, differentiable, smooth, continuous, etc.), so that the response is locally linear, given small enough variation in the input. We assume here that the analyst has control over the design of this experiment. These questions arise:

- What is the operating point?
- How many observations are needed?
- What kinds of variation in the input need to be considered?
- What is the best way to specify the design matrix **X**?

For example, suppose that the input space has dimension two and that we are interested in the effect of ± 10% variation in the input values. Using ± 2% increments on the variables gives a set of 11 values for each input dimension, namely, { -10%, -8%, -6%, -4%, -2%, 0%, 2%, 4%, 6%, 8%, 10% }. Call this set of values **S**. Constructing all possible pairs of the values gives a total of $11^2 = 121$ points. The design is the product set $\mathbf{S} \times \mathbf{S}$, and a design point is a pair of numbers $\mathbf{x}=(x_1, x_2)$. The design matrix **X** has 121 rows, each consisting of a distinct pair **x** from $\mathbf{S} \times \mathbf{S}$. In two dimensions, we can graph the design:

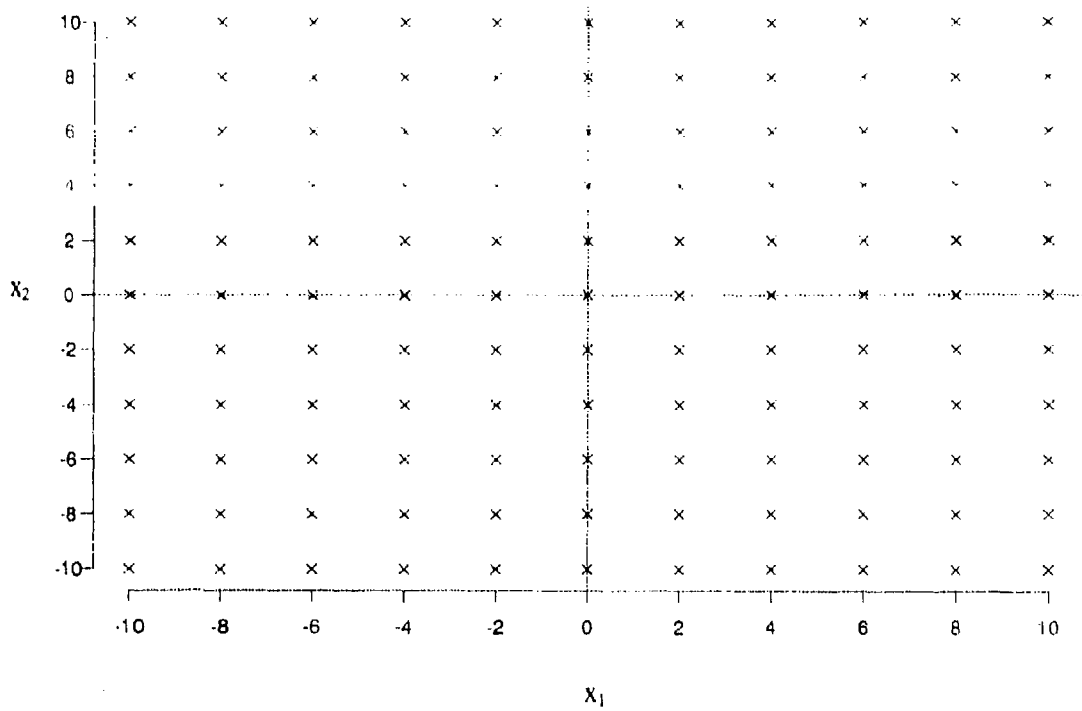


Figure 1. Full-Grid Design

This scheme exercises all possible combinations of the inputs and thus provides good coverage of the input space at the expense of a large number of design points. Note, that in five dimensions, this design requires $11^5 = 161051$ points.

An alternative is to vary only one dimension of **x** at a time, leaving the others fixed at the operating point. With the same set of values **S** as above, the number of points required here is $1+10p$, where p is the dimension of **x** . Here we gain information only along the x coordinate axes and not in the off-axis regions. Predictions derived from such a model will, in general, only be valid when one of the quantities varies $\pm$ 10% and the other is fixed at zero. Such a design is unacceptable if one wishes to make predictions based on both quantities having a variation in the $\pm$ 10% range simultaneously.

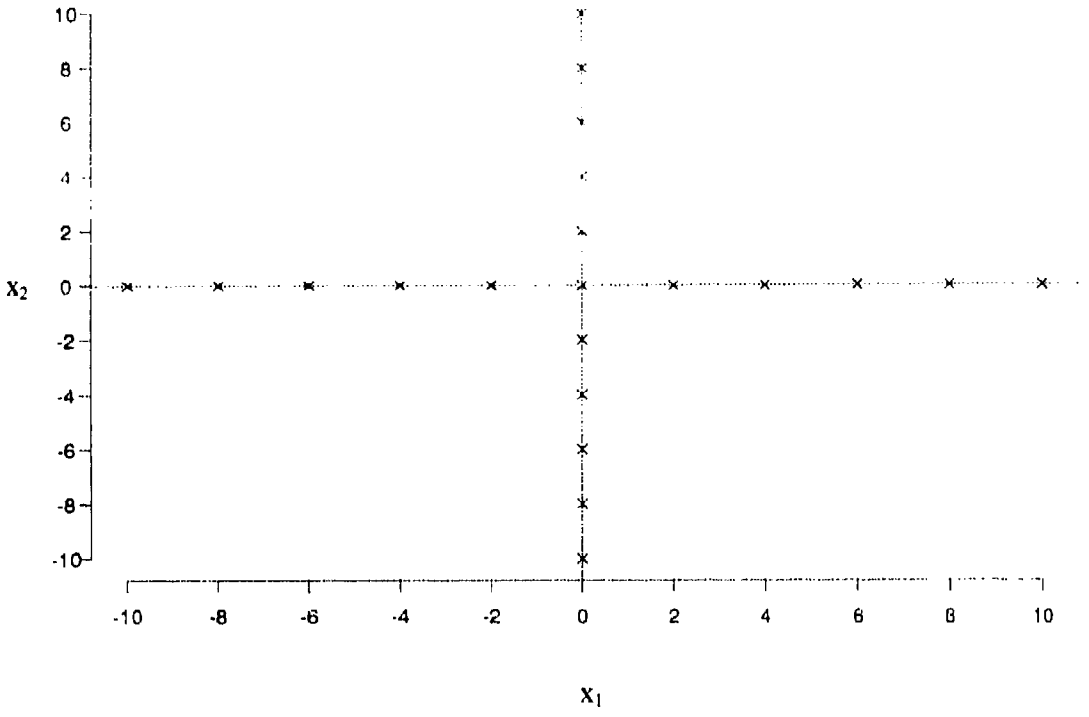


Figure 2. Axis-Only Design

A design which provides complete coverage of the input space and also offers control over the number of points is more useful than the full-grid and axis-only alternatives. The Latin Hypercube Sampling (LHS) design[3] has these desirable characteristics. Consider pairing two random permutations of the base set S to generate a design with 11 points. This procedure is the basis of LHS.

Table 1. LHS with 11 Points in 2 Dimensions

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
X_1	2	0	-6	4	-2	10	8	-8	-10	6	-4
X_2	-8	6	-10	-2	0	2	-6	-4	4	10	8

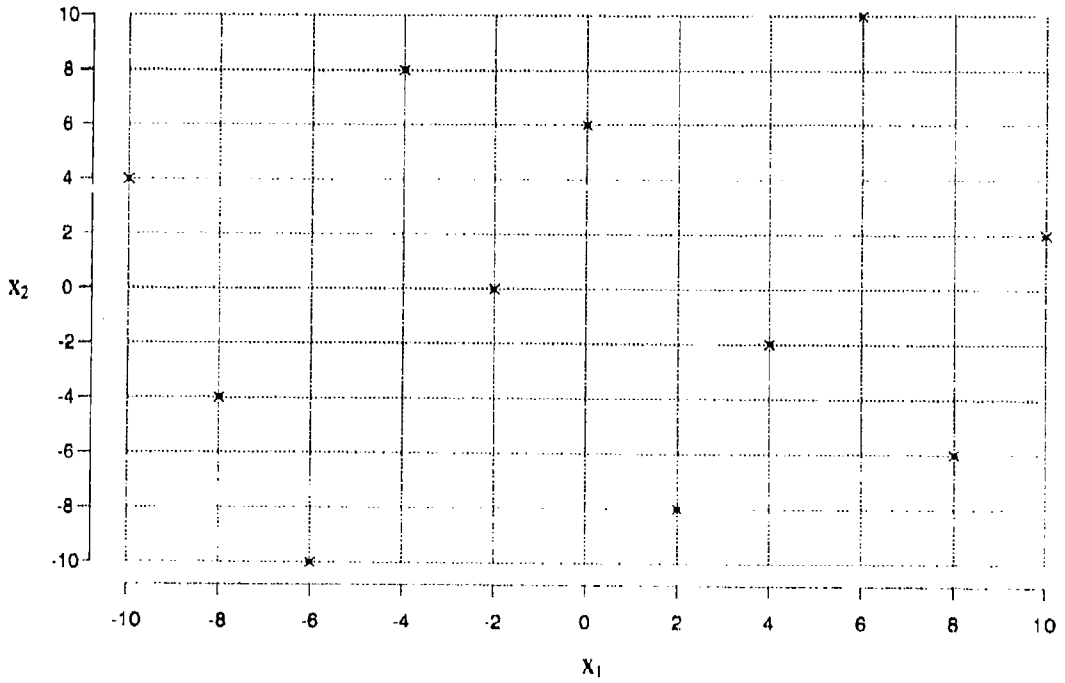


Figure 3. LHS with 11 Points in 2 Dimensions

2. The Latin Hypercube Sample

In full generality, the Latin Hypercube Sample allows free choice of the number of design points (henceforth denoted by n), the dimension of the input space (p), the marginal probability distribution of each of the p input variables, and the correlation structure of the input space.

2.1. Definition

A p-dimensional Latin Hypercube Sample of size n is formed as follows:[4]

> Divide the range of each variable into n bins based on equal width or equal probability. For each variable, select a point at random from each bin. Then randomly order the points for each variable and combine them to form p-tuples. The resulting collection of n vectors, each of length p, is a Latin Hypercube Sample.

We begin by examining the simplest case and proceed to develop generality by presenting examples of the more involved constructions.

2.2. Discrete Uniform Distribution

The simplest case is a one-dimensional (1-D) LHS of size n drawn from the discrete uniform distribution. Without loss of generality, we can take the allowable variable values to be the set $\mathbf{N} = \{ 0, 1, 2, ..., n-1 \}$. The LHS is then a random permutation of **N**. For example, with $n = 10$ we have:

Table 2. Discrete, Uniform, 1-D LHS

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
X	7	0	2	3	6	1	8	9	5	4

Independent 1-D samples are combined to form higher-dimensional samples:

Table 3. Discrete, Uniform, 2-D LHS

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
X_1	7	0	2	3	6	1	8	9	5	4
X_2	2	0	1	9	3	7	5	4	8	6

In the LHS, there is no connection between the number of sample points and the dimension of the input space, in contrast with the previously considered designs.

2.3. Continuous Uniform Distribution

Now we consider a sample with uniform distribution on the unit interval, U(0,1). Construction of this sample is based on the discrete sample of the previous section. For $p = 1$, the 1-D case, an LHS of size n is constructed as follows:

a. Generate a random permutation of the set **N** = { 0, 1, 2, ..., n-1 }.

b. Add a U(0,1) random quantity to each element of **N**.

c. Divide by n to scale the sample into the interval (0,1).

For example, with n=10, we have the results in Table 4.

Table 4. Development of Continuous, Uniform, 1-D LHS

step	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
a.	7	0	2	3	6	1	8	9	5	4
b.	7.690	0.526	2.364	3.884	6.558	1.731	8.667	9.654	5.816	4.272
c.	0.769	0.053	0.236	0.388	0.656	0.173	0.867	0.965	0.582	0.427

Note that we effectively divide the allowable variable range (0,1) into n equiprobable bins (0,0.1), (0.1,0.2), ..., (0.9,1); order the bins randomly; and then select a point from each bin, again with equal probability.

A 2-D LHS is formed by generating independent 1-D samples for each variable. Adding another dimension to the previous example gives the results in Table 5.

Table 5. Continuous, Uniform, 2-D LHS

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
X_1	0.769	0.053	0.236	0.388	0.656	0.173	0.867	0.965	0.582	0.427
X_2	0.215	0.041	0.115	0.929	0.362	0.746	0.560	0.483	0.868	0.609

Higher-dimensional samples are formed by generating independent samples in each dimension.

2.4. Arbitrary Distributions

The LHS examples generated previously have the U(0,1) uniform distribution in each dimension. Transformation to other continuous distributions can be accomplished independently in each dimension by applying the appropriate inverse probability integral transform (inverse cumulative distribution function). The argument is reproduced here:

> Let the random variable U have the uniform distribution on the unit interval. The cumulative distribution function (cdf) of U is $\text{Prob}\{U \le t\} = t$ for $0 \le t \le 1$. We wish to transform U into a random quantity X with a specified cdf F(x). So take $X = F^{-1}(U)$. Then $\text{Prob}\{X \le x\} = \text{Prob}\{F^{-1}(U) \le x\} = \text{Prob}\{U \le F(x)\} = F(x)$ as desired.

2.5. Orthogonal Design in Linear Regression

The design matrix **X** in a linear regression problem (equation [5]) is said to be *orthogonal* if the product **X'X** is a diagonal matrix. The variables (columns) of such a design are then

uncorrelated. In the statistical literature, the term *multicollinearity* refers to a departure from orthogonality. On one hand, orthogonality is an absolute. Either a matrix is orthogonal or it isn't. In contrast, use of the word *multicollinearity* is intended to suggest some degree of linear dependence among a set of vectors. Thus, multicollinearity is subject to quantification and comparison. Common measures of multicollinearity include variance inflation factors, the determinant, various types of matrix metrics, and various definitions of the condition number. Further discussion of multicollinearity is deferred to section 2.7, where several of these measures are defined and used.

One of the computational benefits of an orthogonal design is that calculation of the parameter estimate for one of the variables involves only that particular column of the design matrix (along with the dependent variable), so variables can be added or deleted form the design scheme without recalculating all estimates. Likewise, a single column can be changed and the corresponding parameter can be re-estimated independently of the others. A second advantage of the orthogonal design is the optimal variance property, which essentially states that parameter estimates have minimum variance when the design is orthogonal.[5] Practically speaking, this corresponds to reduced error estimates.

2.6. Correlation and Correlation Conditioning

The rank correlation of any continuous LHS, whether it be drawn from the uniform distribution or an arbitrary distribution, is identically equal to the rank correlation of the underlying discrete uniform LHS from which the sample was obtained. So, all inquiries concerning the rank correlation of an LHS can be addressed by considering the discrete uniform case.

Ideally, the variables (*i.e.*, columns) of an LHS should be uncorrelated, as they were generated independently. In practice, of course, these vectors exhibit nonzero correlation. For example, here is an integer LHS with 10 observations and 5 variables:

$$\mathbf{X} = \begin{bmatrix} 8 & 4 & 8 & 7 & 1 \\ 4 & 5 & 5 & 3 & 2 \\ 5 & 10 & 10 & 8 & 8 \\ 1 & 7 & 9 & 10 & 9 \\ 6 & 2 & 7 & 4 & 7 \\ 9 & 9 & 1 & 5 & 10 \\ 2 & 1 & 6 & 1 & 6 \\ 10 & 3 & 3 & 2 & 5 \\ 3 & 6 & 2 & 9 & 4 \\ 7 & 8 & 4 & 6 & 3 \end{bmatrix}. \qquad [6]$$

To an accuracy of two decimal places, this sample has rank correlation:

$$\rho_X = \begin{bmatrix} 1.00 & 0.07 & -0.39 & -0.30 & -0.13 \\ . & 1.00 & -0.01 & 0.64 & 0.33 \\ . & . & 1.00 & 0.31 & 0.08 \\ . & . & . & 1.00 & 0.15 \\ . & . & . & . & 1.00 \end{bmatrix}. \qquad [7]$$

The ideal correlation structure of such a sample should be the identity matrix **I** (*i.e.*, distinct variables should be uncorrelated). In the 1-D case, we divide each element by the standard deviation (square root of the variance) of the sample to scale the sample variance to unity. Analogous procedures can be used in higher dimensions to produce uncorrelated vectors.[6] One way of accomplishing this "decoupling" in the multidimensional case is presented here.

Let **S** be the sample variance-covariance matrix of **X**. The diagonal elements of **S** are the sample variances of the input vectors, and the off-diagonal elements are the sample covariances. Let **T'T** be the Cholesky decomposition[7] of **S**. Then **T** is upper-triangular and **T'T = S**. Let $\mathbf{Q} = \mathbf{T}^{-1}$ and consider the quantity **XQ**. Applying standard identities[8] concerning the variance of multivariate random quantities, we have

$$\begin{aligned} \text{var}(\mathbf{XQ}) &= \mathbf{Q}' \cdot \text{var}(\mathbf{X}) \cdot \mathbf{Q} \\ &= \mathbf{Q}' \cdot \mathbf{S} \cdot \mathbf{Q} \\ &= \mathbf{Q}' \cdot \mathbf{T}'\mathbf{T} \cdot \mathbf{Q} \\ &= (\mathbf{TQ})'\mathbf{TQ} \\ &= \mathbf{I}'\mathbf{I} \\ &= \mathbf{I} \end{aligned} \qquad [8]$$

so the product **XQ** has unit variance. We have "divided" **X** by the "square root" of its variance to produce an object with the required variance **I**. The resulting correlation structure will also be **I**. In this case,

$$\mathbf{S} = \mathbf{T}'\mathbf{T} = \begin{bmatrix} 9.167 & 0.611 & -3.611 & -2.722 & -1.167 \\ 0.611 & 9.167 & -0.056 & 5.833 & 3.056 \\ -3.611 & -0.056 & 9.167 & 2.833 & 0.722 \\ -2.722 & 5.833 & 2.833 & 9.167 & 1.389 \\ -1.167 & 3.056 & 0.722 & 1.389 & 9.167 \end{bmatrix}, \qquad [9]$$

$$\mathbf{T} = \begin{bmatrix} 3.028 & 0.202 & -1.193 & -0.899 & -0.385 \\ 0 & 3.021 & 0.061 & 1.991 & 1.037 \\ 0 & 0 & 2.782 & 0.589 & 0.072 \\ 0 & 0 & 0 & 2.012 & -0.529 \\ 0 & 0 & 0 & 0 & 2.767 \end{bmatrix}, \qquad [10]$$

$$\mathbf{Q} = \mathbf{T}^{-1} = \begin{bmatrix} 0.330 & -0.022 & 0.142 & 0.128 & 0.075 \\ 0 & 0.331 & -0.007 & -0.325 & -0.186 \\ 0 & 0 & 0.359 & -0.105 & -0.029 \\ 0 & 0 & 0 & 0.497 & 0.095 \\ 0 & 0 & 0 & 0 & 0.361 \end{bmatrix}, \quad [11]$$

and

$$\mathbf{XQ} = \begin{bmatrix} 2.642 & 1.148 & 3.983 & 2.359 & 0.647 \\ 1.321 & 1.567 & 2.329 & -0.151 & 0.230 \\ 1.651 & 3.200 & 4.232 & 0.309 & 1.871 \\ 0.330 & 2.295 & 3.326 & 1.873 & 2.710 \\ 1.982 & 0.530 & 3.354 & 1.368 & 2.782 \\ 2.973 & 2.781 & 1.573 & 0.601 & 3.060 \\ 0.661 & 0.287 & 2.433 & -0.204 & 2.051 \\ 3.303 & 0.772 & 2.477 & 0.981 & 2.101 \\ 0.991 & 1.920 & 1.101 & 2.694 & 1.351 \\ 2.312 & 2.494 & 2.374 & 0.853 & 0.573 \end{bmatrix}. \quad [12]$$

Now let each column of the matrix **Y** contain the ranks of the data in the corresponding column of **XQ**. Transforming data to ranks changes variance but not rank correlation. This operation yields an integer matrix, each column of which can be viewed as a permutation of the corresponding column of the original matrix **X**. The result is:

$$\mathbf{Y} = \begin{bmatrix} 8 & 4 & 9 & 9 & 3 \\ 4 & 5 & 3 & 2 & 1 \\ 5 & 10 & 10 & 3 & 5 \\ 1 & 7 & 7 & 8 & 8 \\ 6 & 2 & 8 & 7 & 9 \\ 9 & 9 & 2 & 4 & 10 \\ 2 & 1 & 5 & 1 & 6 \\ 10 & 3 & 6 & 6 & 7 \\ 3 & 6 & 1 & 10 & 4 \\ 7 & 8 & 4 & 5 & 2 \end{bmatrix} \quad [13]$$

which now has rank correlation

$$\mathbf{Q_Y} = \begin{bmatrix} 1.00 & 0.07 & 0.07 & 0.07 & -0.13 \\ . & 1.00 & -0.09 & -0.02 & -0.05 \\ . & . & 1.00 & 0.07 & 0.12 \\ . & . & . & 1.00 & 0.08 \\ . & . & . & . & 1.00 \end{bmatrix}. \quad [14]$$

Compare this with the rank correlation of the original sample **X** (equation [7]). The effect of such a transformation is not entirely obvious, as most observers are not able to visualize

higher-dimensional objects. Certainly some of the offensive correlations have decreased, but several of the off-diagonal elements of the correlation structure have increased in magnitude. lowever, there are a number of ways to measure multicollinearity in the sample.

2.7. Measures of Performance

Perhaps the most common scalar measures of multicollinearity assiciated with a design matrix are the condition number and determinant measure.

The eigenvalues λ_i of the ideal correlation structure are all equal to 1. The determinant of a matrix is the product of its eigenvalues. Hence, the determinant δ of the ideal structure is also equal to 1. A condition number κ may be defined as the ratio of largest to smallest eigenvalues λ_1/λ_5. This quantity will also be 1 in the ideal case. Finally, we may consider the L^2 norm ε (also called the Euclidean norm or root mean square distance) between the sample's correlation structure and the ideal identity matrix. This quantity should be zero. Refer to the first two lines of Table 6 for measures associated with the samples **X** and **Y**.

By all indications, this procedure has improved (decreased) the correlation of the sample. Note that the product **XQ** indeed has exact unit correlation and that the final step of replacing columns of **XQ** by column ranks again disturbs the correlation structure. It is natural to consider iterative application of the procedure in hopes of obtaining a "limiting" sample with the "most ideal" correlation under the constraint of replacing columns with column ranks. We can repeat the procedure and generate a sample $\mathbf{Y}_+$ from **Y** in the same manner that **Y** was generated from **X**. The details are not reproduced here, but another application of the procedure permutes four elements in last column, and the result is

$$\mathbf{Y}_+ = \begin{bmatrix} 8 & 4 & 9 & 9 & 2 \\ 4 & 5 & 3 & 2 & 1 \\ 5 & 10 & 10 & 3 & 5 \\ 1 & 7 & 7 & 8 & 8 \\ 6 & 2 & 8 & 7 & 9 \\ 9 & 9 & 2 & 4 & 10 \\ 2 & 1 & 5 & 1 & 7 \\ 10 & 3 & 6 & 6 & 6 \\ 3 & 6 & 1 & 10 & 4 \\ 7 & 8 & 4 & 5 & 3 \end{bmatrix}. \qquad [15]$$

The resulting correlation structure is

$$\mathbf{Q}_{\mathbf{Y}_+} = \begin{bmatrix} 1.00 & 0.07 & 0.07 & 0.07 & 0.02 \\ . & 1.00 & -0.09 & -0.02 & -0.03 \\ . & . & 1.00 & 0.07 & 0.04 \\ . & . & . & 1.00 & -0.03 \\ . & . & . & . & 1.00 \end{bmatrix}. \qquad [16]$$

The associated measures of multicollinearity are presented in Table 6.

Table 6. Measures of Multicollinearity

sample	λ_1	λ_2	λ_3	λ_4	λ_5	κ	δ	ε
X	1.960	1.335	0.879	0.601	0.224	8.751	0.310	1.810
Y	1.275	1.077	0.943	0.877	0.828	1.541	0.940	0.130
$\mathbf{Y}_+$	1.147	1.085	1.015	0.905	0.849	1.350	0.970	0.061
ideal	1	1	1	1	1	1	1	0

For this particular sample, the process has terminated. Another application does not change the sample.

2.8. The General Effect of Correlation Conditioning

We can demonstrate the effect of correlation correction by considering this simulation:

- Generate 1,000 LH samples, each with 100 observations in 5 variables, with no correlation correction.
- Generate 1,000 LH samples, each with 100 observations in 5 variables, with a single step of correlation correction.
- Generate 1,000 LH samples, each with 100 observations in 5 variables, with completed correlation correction.
- Compute and compare the cumulative probability distributions of
  κ, the condition numbers,
  δ, the determinants of the empirical correlation structures $\hat{\mathbf{\varrho}}$, and
  ε, the L^2 norms $|\hat{\mathbf{\varrho}}-\mathbf{I}|_2$
  for each of the three sets of samples.

Figures 4 through 6 depict the empirical distributions of κ, δ, and ε. Neither the single example nor the simulation provides proof of the effectiveness of the correlation correction procedure, but the indication is that single correction substantially improves the behavior of the sample, and that completed correction further improves the behavior of the sample.

Table 7 details results from the simulation in the form of empirical quantiles (q) for each measure of multicollinearity (κ, δ, and ε) at each level of correlation correction (none, single, and complete). Such tabulations facilitate quantitative observations about the distributions under study. For example:

> Note that 99% of κ_0 lies above 1.349, whereas 99% of κ_1 lies below 1.127. This is complete separation of distributions, for all practical purposes. Also, 85% of κ_1 lies above 1.046, whereas 99% of κ_{∞} lies below 1.046. Empirically, this indicates a

probability of 0.85 that the condition number of an LHS with single correlation correction exceeds 1.046, and a probability of 0.01 that the condition number of an LHS with completed correlation correction does not exceed 1.046. This may be an important difference in practical applications.

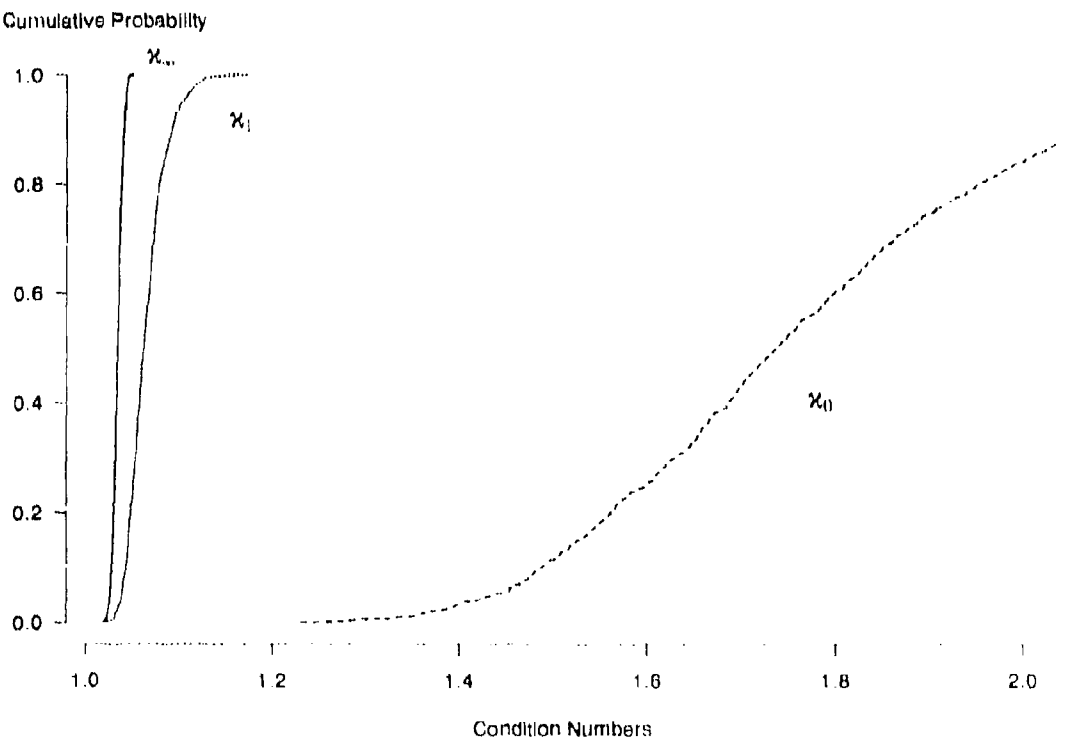


Figure 4. Condition Numbers

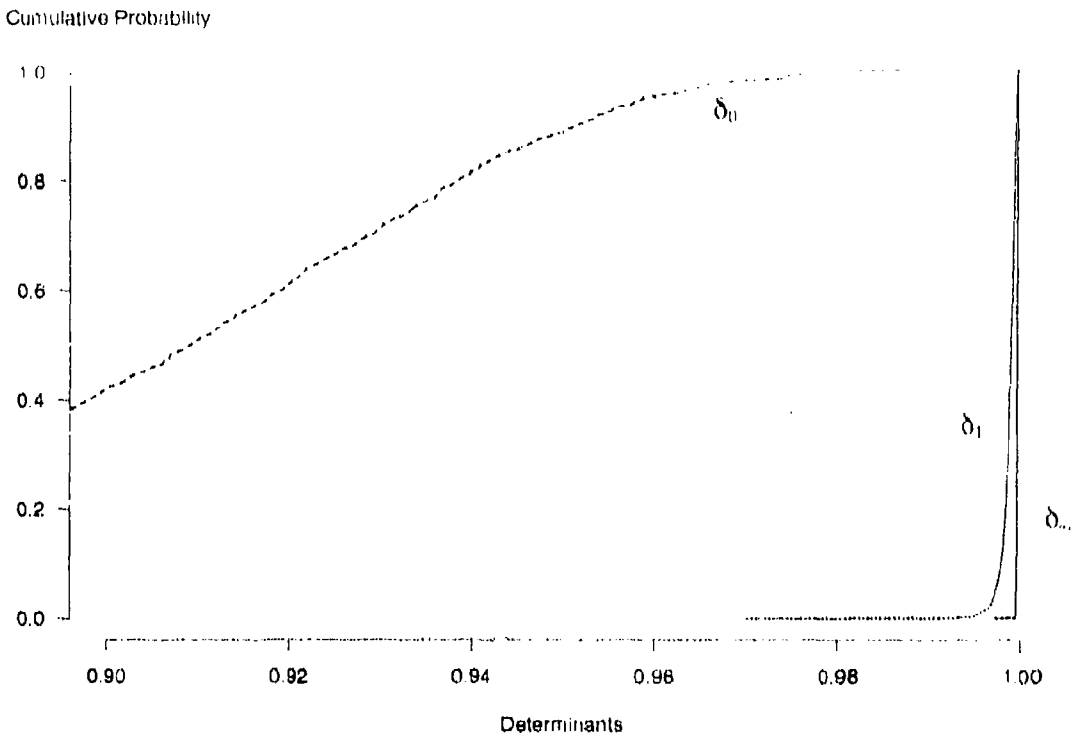


Figure 5. Determinants

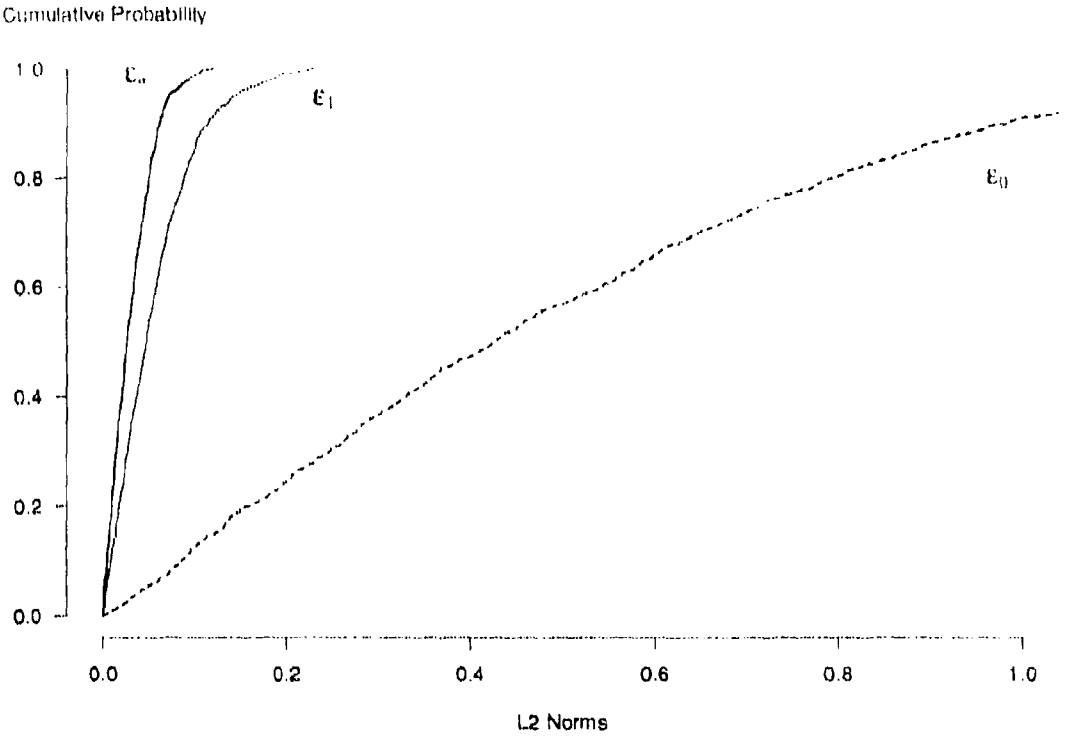


Figure 6. L^2 Norms

Table 7. Quantiles of Performance Measures

q	Condition Number ($\varkappa$)			Determinant (δ)			L^2 Norm (ε)		
	$\varkappa_0$	$\varkappa_1$	$\varkappa_\infty$	δ_0	δ_1	δ_∞	ε_0	ε_1	ε_∞
1%	1.349	1.033	1.025	0.7913	0.9956	0.9994	0.011	0.001	0.000
5%	1.435	1.040	1.028	0.8369	0.9971	0.9994	0.048	0.004	0.002
10%	1.489	1.043	1.029	0.8523	0.9977	0.9995	0.086	0.009	0.005
15%	1.531	1.046	1.031	0.8637	0.9979	0.9995	0.122	0.013	0.007
20%	1.563	1.049	1.031	0.8713	0.9982	0.9995	0.158	0.017	0.010
25%	1.601	1.051	1.032	0.8784	0.9983	0.9995	0.205	0.022	0.012
30%	1.630	1.053	1.033	0.8851	0.9985	0.9996	0.247	0.026	0.014
35%	1.658	1.056	1.034	0.8927	0.9986	0.9996	0.284	0.030	0.017
40%	1.687	1.058	1.034	0.8981	0.9987	0.9996	0.330	0.036	0.021
45%	1.710	1.060	1.035	0.9039	0.9987	0.9996	0.370	0.042	0.023
50%	1.739	1.063	1.036	0.9092	0.9988	0.9996	0.427	0.047	0.025
55%	1.766	1.065	1.036	0.9139	0.9989	0.9996	0.473	0.052	0.029
60%	1.801	1.069	1.037	0.9190	0.9990	0.9996	0.545	0.059	0.032
65%	1.836	1.071	1.037	0.9231	0.9991	0.9996	0.597	0.064	0.036
70%	1.866	1.074	1.038	0.9286	0.9991	0.9997	0.657	0.070	0.041
75%	1.909	1.077	1.039	0.9335	0.9992	0.9997	0.721	0.078	0.045
80%	1.956	1.080	1.040	0.9384	0.9993	0.9997	0.801	0.087	0.050
85%	2.014	1.087	1.041	0.9437	0.9994	0.9997	0.891	0.099	0.056
90%	2.084	1.094	1.042	0.9511	0.9994	0.9997	0.992	0.115	0.062
95%	2.183	1.103	1.044	0.9590	0.9995	0.9998	1.178	0.141	0.072
99%	2.445	1.127	1.046	0.9745	0.9997	0.9998	1.640	0.196	0.101

key 0 : no correlation correction
1 : single correlation correction
∞ : completed correlation correction

2.9. Inducing Correlation in the Sample

We can modify the procedure outlined in section 2.6 to induce a desired correlation in the Latin Hypercube Sample.

Again, we begin with an integer LHS **X** having sample variance-covariance matrix **S**. As before, **T**'**T** is the Cholesky decomposition of **S**, and $\mathbf{Q} = \mathbf{T}^{-1}$. Now suppose that the desired correlation structure of the sample is **C**. Let **R**'**R** be the Cholesky decomposition of **C**, and consider the product **XQR**.

$$\begin{aligned} \text{var}(\mathbf{XQR}) &= (\mathbf{QR})' \cdot \text{var}(\mathbf{X}) \cdot \mathbf{QR} \\ &= \mathbf{R}'\mathbf{Q}' \cdot \mathbf{S} \cdot \mathbf{QR} \\ &= \mathbf{R}'\mathbf{Q}' \cdot \mathbf{T}'\mathbf{T} \cdot \mathbf{QR} \\ &= \mathbf{R}' \cdot (\mathbf{TQ})'\mathbf{TQ} \cdot \mathbf{R} \\ &= \mathbf{R}' \cdot \mathbf{I}'\mathbf{I} \cdot \mathbf{R} \\ &= \mathbf{R}'\mathbf{R} \\ &= \mathbf{C}. \end{aligned} \qquad [17]$$

Now **XQR** has variance (and hence correlation) exactly equal to **C**. Let each column of the matrix **Y** contain the ranks of the data in the corresponding column of **XQR**. Continue as before, treating **Y** as the new sample and iterating the procedure. The benefits observed in the unit-correlation case (section 2.8) carry through to the arbitrary-correlation case.

3. Some Uses of the Latin Hypercube Sample

3.1. Local Sensitivity Analysis: Single Operating Point

The most basic use of LHS is to model small variation in input around a single operating point using an uncorrelated uniform sample. This is described in detail in the Introduction of this report and illustrated in Figure 3.

3.2. Local Sensitivity Analysis: Multiple Operating Points

Suppose now that we are interested in the local sensitivity of our model at a number of operating points. For the sake of illustration, take p=2. Consider three operating points, say, $x_1 = (1, 2)$, $x_2 = (2, 1)$, and $x_3 = (4, 3)$. We impose $\pm$ 10% variation on the inputs and generate 25 perturbations of each operating point. Initial operating points are indicated by "+" and LHS points by "·" in Figure 7.

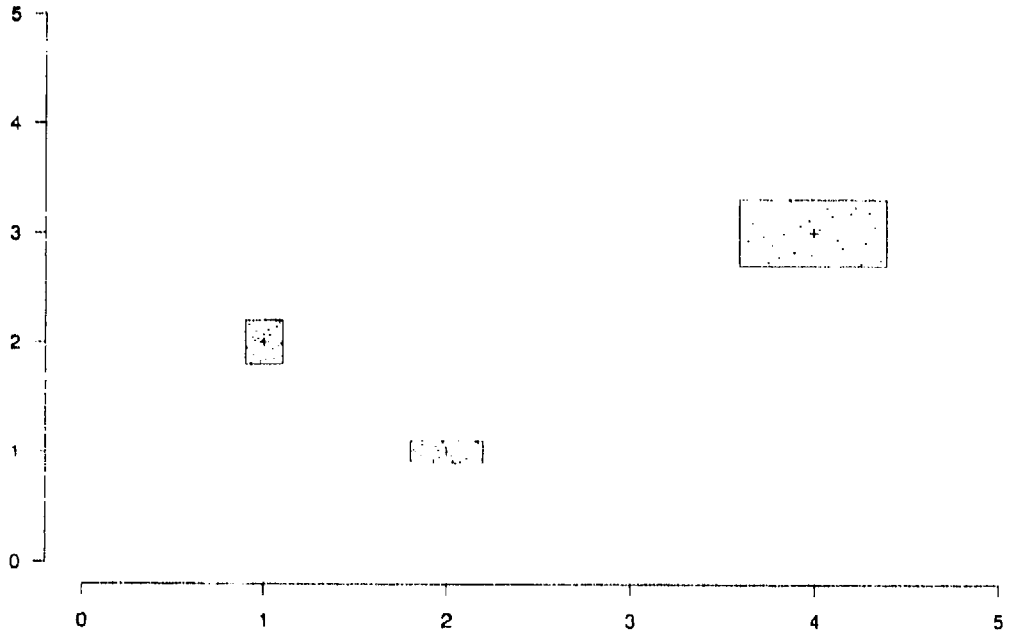


Figure 7. Multiple Local Operating Points

3.3. Global Sensitivity Analysis

We may desire to "connect" the space between operating points and develop a global model of the simulation under study. In this case, it may be appropriate to induce a particular correla-

tion in the LHS to provide sampling in desirable regions. Refer to Figure 8. Initial operating points are indicated by "+" and LHS points by ".". Suppose we choose to regard the points (1,5) and (5,1) as infeasible and the points (1,1) and (5,5) as realistic extensions of the operating space. The sample in Figure 8 was generated with a correlation of 0.7 between dimensions, and it apparently conforms to this notion of feasibility. The uniform distribution was used here, but other distributions may be appropriate depending on the application. Note that changing marginal distributions through use of the inverse probability integral transform does not change the rank correlation of the sample, as the mapping is monotonic increasing.

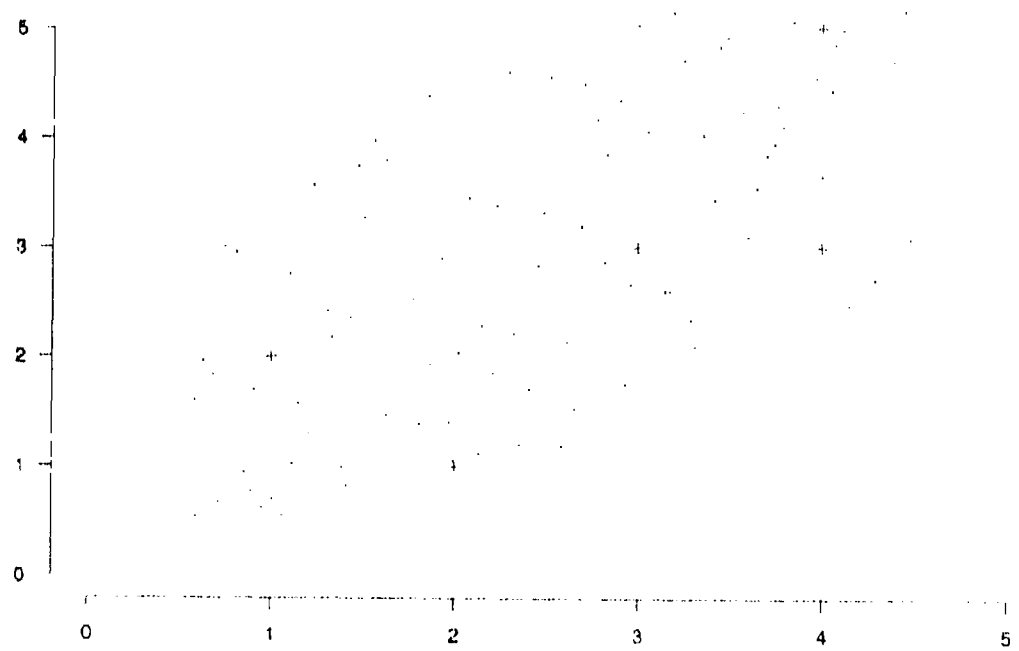


Figure 8. Global Analysis

4. Conclusions and Recommendations

The Latin Hypercube Sample is appropriately used to generate input for simulation model sensitivity analyses.[9] Consider the linear regression problem (equation [5]) developed in the introduction of this report. It is well known[10] that the variance of the parameter estimate increases as the correlation of the input variables increases, the ideal (minimum variance) case being that of uncorrelated inputs. Various ways of quantifying multicollinearity, or departure from orthogonality, have been suggested. These including the condition number and determinant measures. Statisticians agree that a design with minimal correlation among the input variables is desirable. However, as Stuart and Ord[11] point out, the word *minimal* in this context does not have a unique interpretation:

> Stewart ... presents several indices for assessing multicollinearity; the ensuing discussion indicates the lively debate that persists.

Correlation correction in Latin Hypercube Sampling reduces popular measures of multicollinearity. This increases the efficiency of subsequent statistical procedures. Therefore, the correction should be applied when efficiency is an issue.

The effects of higher correlation are amplified when dimensionality of the sample is high and the number of points in the sample (cardinality of the sample) is low. Schemes which use a large number of high-dimensional, low-cardinality samples may particularly benefit from completed correlation correction.

5. References

[1] Ronald L. Iman, John C. Helton, and James E. Campbell, "An Approach to Sensitivity Analysis of Computer Models: Part I – Introduction, Input Variable Selection and Preliminary Variable Assessment," Journal of Quality Technology, Vol. 13, No. 3, July 1981.

[2] Phillip J. Hanes, Scott L. Henry, Gary S. Moss, Karen R. Murray, and Wendy A. Winner, "Modular UNIX®-based Vulnerability Estimation Suite (MUVES) Analyst's Guide," Ballistic Research Laboratory Memorandum Report No. 3954, December 1991.

[3] Iman, Helton, and Campbell, *idem.*

[4] Iman, Helton, and Campbell, *idem.*

[5] G. A. F. Seber, Linear Regression Analysis, John Wiley & Sons, 1977.

[6] Ronald L. Iman and W. J. Conover, "A Distribution-Free Approach to Inducing Rank Correlation Among Input Vectors," Communications in Statistics – Simulation and Computing, Vol. 11, No. 3, 1982, pp. 311-334.

[7] Ronald A. Thisted, Elements of Statistical Computing – Numerical Computation, Chapman & Hall, 1988.

[8] Richard A. Johnson and Dean W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, 1988.

[9] M. D. McKay, R. J. Beckman, and W. J. Conover, "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," Technometrics, Vol. 21, No. 2, May 1979, pp. 239-245.

[10] Raymond H. Myers, Classical and Modern Regression with Applications, PWS-KENT, 1989.

[11] Alan Stuart and J. Keith Ord, Kendall's Advanced Theory of Statistics, Oxford University Press, 1991.

Intentionally Left Blank.

No. of
Copies Organization

2 Administrator
Defense Technical Info Center
ATTN: DTIC-DDA
Cameron Station
Alexandria, VA 22304-6145

1 Commander
U.S. Army Materiel Command
ATTN: AMCAM
5001 Eisenhower Ave.
Alexandria, VA 22333-0001

1 Director
U.S. Army Research Laboratory
ATTN: AMSRL-OP-SD-TA,
Records Management
2800 Powder Mill Rd.
Adelphi, MD 20783-1145

3 Director
U.S. Army Research Laboratory
ATTN: AMSRL-OP-SD-TL,
Technical Library
2800 Powder Mill Rd.
Adelphi, MD 20783-1145

1 Director
U.S. Army Research Laboratory
ATTN: AMSRL-OP-SD-TP,
Technical Publishing Branch
2800 Powder Mill Rd.
Adelphi, MD 20783-1145

2 Commander
U.S. Army Armament Research,
Development, and Engineering Center
ATTN: SMCAR-TDC
Picatinny Arsenal, NJ 07806-5000

1 Director
Benet Weapons Laboratory
U.S. Army Armament Research,
Development, and Engineering Center
ATTN: SMCAR-CCB-TL
Watervliet, NY 12189-4050

1 Director
U.S. Army Advanced Systems Research
and Analysis Office (ATCOM)
ATTN: AMSAT-R-NR, M/S 219-1
Ames Research Center
Moffett Field, CA 94035-1000

No. of
Copies Organization

1 Commander
U.S. Army Missile Command
ATTN: AMSMI-RD-CS-R (DOC)
Redstone Arsenal, AL 35898-5010

1 Commander
U.S. Army Tank-Automotive Command
ATTN: AMSTA-JSK (Armor Eng. Br.)
Warren, MI 48397-5000

1 Director
U.S. Army TRADOC Analysis Command
ATTN: ATRC-WSR
White Sands Missile Range, NM 88002-5502

1 Commandant
U.S. Army Infantry School
ATTN: ATSH-WCB-O
Fort Benning, GA 31905-5000

Aberdeen Proving Ground

2 Dir, USAMSAA
ATTN: AMXSY-D
AMXSY-MP, H. Cohen

1 Cdr, USATECOM
ATTN: AMSTE-TC

1 Dir, USAERDEC
ATTN: SCBRD-RT

1 Cdr, USACBDCOM
ATTN: AMSCB-CII

1 Dir, USARL
ATTN: AMSRL-SL-I

5 Dir, USARL
ATTN: AMSRL-OP-AP-L

No of
Copies Organization

2 Superintendent
U.S. Military Academy
ATTN: MAIN-SC-A
COL Rickey A. Kolb
Sue M. Bass
West Point, NY 10996-5000

Aberdeen Proving Ground

1 Dir USAMSAA
ATTN: AMXSY-RA,
L. Wald

3 Dir, USAMSAA
ATTN: AMXSY-CD,
D. Nuzman
W. Yeakel
F. Wofford

3 Cmdr, USAARDEC
ATTN: SMCAR-FSF-T,
M. Andriolo
R. Carter
B. Laurie

1 Dir, USARL
ATTN: AMSRL-CI-CA,
A. Celmins

2 Dir, USARL
ATTN: AMSRL-CI-CC,
A. Brodeen
R. Kaste

2 Dir, USARL
ATTN: AMSRL-CI-S,
M. Taylor
B. Bodt

1 Dir, USARL
ATTN: AMSRL-HR-MB,
R. Tauson

1 Dir, USARL
ATTN: AMSRL-HR-SD,
J. Grynovicki

1 Dir, USARL
ATTN: AMSRL-SL-C,
W. Hughes (E333)

1 Dir, USARL
ATTN: AMSRL-SL-B,
P. Deitz (328)

No of
Copies Organization

Aberdeen Proving Ground continued

2 Dir, USARL
ATTN: AMSRL-SL-BA,
J. Walbert (1068)
L. Roach (1068)

3 Dir, USARL
ATTN: AMSRL-SL-BG,
D. Kirk (238)
J. Abell (238)
L. Losie (238)

1 Dir, USARL
ATTN: AMSRL-SL-BL,
D. Bely (328)

1 Dir, USARL
ATTN: AMSRL-SL-BS,
J. Jacobson (328)

14 Dir, USARL
ATTN: AMSRL-SL-BV,
J. Smith (247)
W. Baker (247)
J. Collins (247) (5 copies)
K. Murray (247)
L. Moss (247)
J. Ploskonka (247)
R. Saucier (247)
R. Shnidman (247)
P. Tanenbaum (247)
W. Winner (247)

1 Dir, USARL
ATTN: AMSRL-SL-I,
M. Starks (433)

1 Dir, USARL
ATTN: AMSRL-WT-PA,
D. Webb

1 Dir, USARL
ATTN: AMSRL-WT-TD,
T. Farrand

1 Dir, USARL
ATTN: AMSRL-WT-WB,
A. Thompson

3 Dir, USARL
ATTN: AMSRL-WT-WE,
J. Temperley
J. Thomas
J. Wald

No of
Copies Organization

Aberdeen Proving Ground continued

2 Cmdr, USACSTA
ATTN: STECS-DA-PS,
V. Visnaw
T. Walker

1 USAMRICD
Wood Technical Library
ATTN: SGRD-UV-AI-W,
R. Lee

1 Cmdr, USATECOM
ATTN: AMSTE-TA-G,
N. Dunn
M. Prather
P. Smyers

1 Cmdr, USATECOM
ATTN: AMSTE-TA-O,
J. Hires

1 Cmdr, USATECOM
ATTN: AMSTE-TA-S,
T. Kocher

1 Dir, USAMSAA
ATTN: AMXSY-GC,
R. Sandmeyer

1 Cmdr, CBDCOM
ATTN: SCBRD-AST,
L. Sturdivan

INTENTIONALLY LEFT BLANK.

USER EVALUATION SHEET/CHANGE OF ADDRESS

This Laboratory undertakes a continuing effort to improve the quality of the reports it publishes. Your comments/answers to the items/questions below will aid us in our efforts.

1. ARL Report Number ARL-TR-586 Date of Report October 1994

2. Date Report Received _____

3. Does this report satisfy a need? (Comment on purpose, related project, or other area of interest for which the report will be used.) _____

4. Specifically, how is the report being used? (Information source, design data, procedure, source of ideas, etc.) _____

5. Has the information in this report led to any quantitative savings as far as man-hours or dollars saved, operating costs avoided, or efficiencies achieved, etc? If so, please elaborate. _____

6. General Comments. What do you think should be changed to improve future reports? (Indicate changes to organization, technical content, format, etc.) _____

CURRENT ADDRESS

_____
Organization

_____
Name

_____
Street or P.O. Box No.

_____
City, State, Zip Code

7. If indicating a Change of Address or Address Correction, please provide the Current or Correct address above and the Old or Incorrect address below.

OLD ADDRESS

_____
Organization

_____
Name

_____
Street or P.O. Box No.

_____
City, State, Zip Code

(Remove this sheet, fold as indicated, tape closed, and mail.)
(DO NOT STAPLE)

DEPARTMENT OF THE ARMY

OFFICIAL BUSINESS

NO POSTAGE
NECESSARY
IF MAILED
IN THE
UNITED STATES

BUSINESS REPLY MAIL
FIRST CLASS PERMIT NO 0001, APG, MD

Postage will be paid by addressee

Director
U.S. Army Research Laboratory
ATTN: AMSRL-OP-AP-L
Aberdeen Proving Ground, MD 21005-5066